Universidade de Brasília

Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Análise de Componentes Principais na avaliação da motilidade gástrica por cintilografia

Ana Carolina da Cruz

Orientador: Prof. Dr. George Freitas von Borries

Brasília

2018

Ana Carolina da Cruz

Análise de Componentes Principais na avaliação da motilidade gástrica por cintilografia

Relatório apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. George Freitas von Borries

Brasília

2018

Ana Carolina da Cruz
Análise de Componentes Principais na avaliação da motilidade gástrica por cintilografia
/ Ana Carolina da Cruz. – Brasília, 2018-
39 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. George Freitas von Borries

Relatório Final – Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística
Trabalho de Conclusão de Curso de Graduação, 2018.

1. Análise de Componentes Principais. 2. Imagens. 3. Multivariada. 4. Motilidade Gástrica. 5. Velocidade da digestão. 6. Limpeza.

Ana Carolina da Cruz

Análise de Componentes Principais na avaliação da motilidade gástrica por cintilografia

Relatório apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Trabalho aprovado. Brasília, 29 de junho de 2018:

Prof. Dr. George Freitas von Borries
Orientador

Prof. Dr. Guilherme Souza Rodrigues
Membro da Banca

Prof. Dr. Bernardo Borba de Andrade
Membro da Banca

Brasília
2018

Resumo

Sintomas em problemas de motilidade gastrointestinal podem ser resultantes de uma acomodação gástrica inadequada. Técnicas para verificar a motilidade gástrica por meio de imagens por cintilografia tem sido limitadas somente a inspeção visual. Este estudo propõe a avaliação automática do esvaziamento gástrico utilizando Análise de Componentes Principais bidimensional. Sendo que com as técnicas utilizadas foi possível limpar as imagens por meio da eliminação de ruídos, identificar a região de interesse de forma automática e avaliar a velocidade da digestão dos indivíduos avaliados.

Palavras-chave: Motilidade gástrica, imagens, limpeza, região de interesse, Análise de Componentes Principais 2D.

Abstract

Symptoms in Gastrointestinal dysmotility disorders could be resulted from inadequate gastric accomodation. Tecnhiques for assessing gastric motility through standard scintigraphic gastric images has been made only through visual inspection. This study proposes an automatic evaluation of gastric emptying using bi-dimensional Principal Components Analysis. With the applied techniques it was possible to clean the images, to automatically identify the region of interest and to evaluate the velocity of digestion of evaluated individuals.

Keywords: Gastric motility, images, cleaning, region of interest, 2D Principal Component Analysis.

Sumário

Introdução

A análise da motilidade gástrica permite entender o comportamento digestivo dos indivíduos, identificando possíveis distúrbios gastrointestinais e auxiliando na rapidez da prescrição de medidas preventivas.

Atualmente, a motilidade gástrica é analisada por meio de um estudo que consiste em avaliar a capacidade do estômago em esvaziar um determinado conteúdo gástrico. O esvaziamento gástrico é usualmente analisado por meio de um exame de cintilografia, em que a digestão de uma refeição radiomarcada de um determinado indivíduo é acompanhada durante um período de quatro horas. Durante esse período são geradas imagens a partir da radiação emitida pela refeição ingerida, obtendo informações sobre o comportamento digestivo.

Um dos problemas da análise de imagens por cintilografia é que as imagens possuem informação de regiões que não são de interesse, sendo necessário então identificar a região de interesse (somente o que representaria o estômago) em cada imagem. Atualmente, esse processo de identificação não é automático. Uma vez identificada a região de interesse, o próximo passo consiste na avaliação do esvaziamento gástrico durante o período avaliado, obtendo informações a respeito da motilidade gástrica do indivíduo, que irão servir de ferramenta para diagnósticos.

Este trabalho utiliza as imagens de motilidade gástrica de seis indivíduos para sugerir técnicas estatísticas que facilitem o processo de análise das imagens, por meio da identificação automática da região de interesse, limpeza das imagens e avaliação da velocidade de digestão utilizando componentes principais.

1 Metodologia

Ao se trabalhar com imagens, usualmente nos deparamos com análise de dados superdimensionados, em que o número de variáveis é proporcionalmente maior que o número de observações. Isto ocorre porque podemos considerar cada pixel na imagem como uma variável a ser analisada e o número de imagens como observações do conjunto de dados, sendo que normalmente o número de pixels é consideravelmente maior que o número de imagens.

Com a crescente quantidade de dados disponível, tornou-se necessário a aplicação e desenvolvimento de novas metodologias aplicáveis a dados superdimensionados, a fim de extrair informações de bases como microarranjos, espectrometria, cintilografia, eletroencefalografia e outras fontes complexas de dados.

Na análise de grandes bases de dados normalmente são aplicadas duas técnicas: Mineração de dados e Aprendizado Estatístico.

A Mineração de dados tem como objetivo extrair informação dos dados por meio da aplicação de técnicas multivariadas, lidando com bases de dados em que a proporção de observações é alta em relação ao número de variáveis, tornando os testes estatísticos tradicionais muitas vezes pouco conservadores.

O Aprendizado Estatístico, por sua vez, pode ser utilizado em casos que a proporção de variáveis em relação ao número de observações é elevada, comprometendo técnicas inferenciais que possuem pressupostos na Teoria Assintótica. O Aprendizado Estatístico se divide em: supervisionado e não-supervisionado. O primeiro necessita de um conjunto de dados de treinamento, sendo que essa informação é utilizada como base para aplicação desse método. O Aprendizado supervisionado é utilizado em diversas técnicas estatísticas como: Classificação, Modelos Lineares, Redes Neurais, Máquinas de Suporte Vetorial, dentre outras. Já o Aprendizado não-supervisionado não possui qualquer informação a priori sobre os dados. As técnicas estatísticas que utilizam esse tipo de aprendizado possuem um foco mais descritivo do que inferencial. O Aprendizado não-supervisionado é utilizado usualmente em Análise de Agrupamento, Análise de Componentes Principais, Escalonamento Multidimensional, dentre outras técnicas estatísticas. É possível encontrar mais detalhes sobre essas técnicas em Theodoridis e Koutroumbas (2006).

A Análise de Componentes Principais é uma das principais técnicas estatísticas utilizadas para análise de dados superdimensionados e mais ainda para análise de imagens. Sendo que neste trabalho, a Análise de Componentes Principais foi utilizada para a análise da motilidade gástrica por cintilografia. Neste Capítulo serão abordadas variações da Análise de Componentes Principais para reconhecimento de padrões em imagens.

1.1 Análise de Componentes Principais

Normalmente na análise de dados superdimensionados, as bases de dados são muito grandes, dificultando o entendimento da estrutura dos dados, principalmente visualmente. Sendo assim, umas das principais técnicas utilizadas na análise de dados superdimensionados é a Análise de Componentes Principais.

A Análise de Componentes Principais (ACP) é uma técnica multivariada, que tem como objetivo explicar a variabilidade de um conjunto de variáveis com menor perda de informação possível, por meio de poucas combinações lineares dessas variáveis. Essas combinações lineares, denominadas componentes possuem algumas propriedades, sendo não correlacionadas entre si e são estimadas seguindo uma ordem determinada pela proporção da variabilidade explicada em relação a variação total presente nos dados. Os objetivos gerais dessa técnica consistem na redução da dimensionalidade e na possível facilidade de interpretação da estrutura dos dados.

Suponha $\mathbf{X}_1, \ldots, \mathbf{X}_p$ em $\mathbb{R}^n$, um conjunto de variáveis aleatórias. Neste caso, é possível obter até p combinações lineares (componentes) dessas variáveis. Usualmente, para explicar toda a variabilidade desse conjunto é necessário utilizar quase todas ou todas as componentes obtidas, porém normalmente é possível explicar grande parte da variabilidade dos dados por meio da utilização de poucas componentes, arbitrariamente k componentes. Sendo possível assim substituir as variáveis originais pelas componentes obtidas, reduzindo o conjunto para k ($k \ll p$) variáveis em vez de p variáveis.

Pelo ponto de vista algébrico, as componentes principais representam a seleção de um novo sistema de coordenadas ortogonal e não correlacionado. As novas coordenadas representam as direções em que a variabilidade é máxima, sendo possível assim obter uma estrutura de covariâncias e variâncias mais simples e parcimoniosa.

A estimação das componentes principais se dá por meio da utilização da matriz de covariâncias ou de correlações das variáveis originais. Suponha que $\mathbf{X}' = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ possui a matriz de covariâncias $\mathbf{\Sigma}$ com autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Sendo que para cada autovalor, existe um autovetor $\mathbf{a}_1, \ldots, \mathbf{a}_p$ associado.

Utilizando os autovetores é possível obter as seguintes combinações lineares, conhecidas como componentes principais:

$$\mathbf{Y}_1 = \mathbf{a}_1'\mathbf{X} = a_{11}\mathbf{X}_1 + \cdots + a_{1p}\mathbf{X}_p \tag{1.1}$$

$$\mathbf{Y}_2 = \mathbf{a}_2'\mathbf{X} = a_{21}\mathbf{X}_1 + \cdots + a_{2p}\mathbf{X}_p \tag{1.2}$$

$$\vdots$$

$$\mathbf{Y}_p = \mathbf{a}_p'\mathbf{X} = a_{p1}\mathbf{X}_1 + \cdots + a_{pp}\mathbf{X}_p \tag{1.3}$$

Sendo que,

$$\text{Var}(\mathbf{Y}_i) = \mathbf{a}_i' \boldsymbol{\Sigma}\ \mathbf{a}_i, \quad i = 1, \ldots, p \tag{1.4}$$

$$\text{Cov}(\mathbf{Y}_i, \mathbf{Y}_k) = \mathbf{a}_i' \boldsymbol{\Sigma}\ \mathbf{a}_k, \quad i \neq k, \;\; i, k = 1, \ldots, p \tag{1.5}$$

Vale ressaltar que $\text{Var}(\mathbf{Y}_1)$ pode ser aumentada ao multiplicarmos $\mathbf{a}_1$ por alguma constante. Devido a isso, utiliza-se os autovetores normalizados. Logo, a primeira componente principal $\mathbf{Y}_1 = \mathbf{a}_1'\mathbf{X}$ é representada pela combinação linear que possui a maior variabilidade entre todas as componentes, ou seja, em que $\text{Var}(\mathbf{Y}_1) = \mathbf{a}_1' \boldsymbol{\Sigma}\ \mathbf{a}_1$ é máxima e $\|\mathbf{a}_1\| = 1$. A segunda componente $\mathbf{Y}_2 = \mathbf{a}_2'\mathbf{X}$, $\|\mathbf{a}_2\| = 1$ é determinada pela combinação linear que possui a maior variabilidade e que não seja correlacionada com a primeira componente $\mathbf{Y}_1$. Seguindo o mesmo raciocínio, temos que a i-ésima componente $\mathbf{Y}_i = \mathbf{a}_i'\mathbf{X}$, $\|\mathbf{a}_i\| = 1$ é dada pela combinação linear que possui variabilidade máxima e que não seja correlacionada com as $i - 1$ componentes obtidas.

Como mencionado, as componentes principais são não correlacionadas entre si, sendo uma das vantagens da Análise de Componentes Principais, pois usualmente a ACP é utilizada como uma etapa intermediária para outras técnicas, como na Análise de Agrupamento e na Regressão Múltipla, sendo que com a utilização das componentes o problema de multicolinearidade é resolvido.

Outras propriedades das componentes são apresentadas abaixo:

- A variância da i-ésima componente principal ($\mathbf{Y}_i$) é dada pelo i-ésimo autovalor de $\boldsymbol{\Sigma}$ (λ_i) :

$$\text{Var}(\mathbf{Y}_i) = \lambda_i \tag{1.6}$$

- A primeira componente apresenta a maior variância entre todas as componentes, sendo possível ordenar as componentes da seguinte maneira:

$$\text{Var}(\mathbf{Y}_1) \geq \text{Var}(\mathbf{Y}_2) \geq \cdots \geq \text{Var}(\mathbf{Y}_p) \tag{1.7}$$

- A variação total dos dados é igual a soma dos autovalores que é igual a soma das variância de todas as componentes principais:

$$\sigma_{11} + \cdots + \sigma_{pp} = \text{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \text{Var}(\mathbf{X}_i) = \lambda_1 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{Var}(\mathbf{Y}_i) \tag{1.8}$$

A contribuição de cada uma das componentes é determinada pela proporção da variabilidade explicada em relação a variação total dos dados:

$$\frac{\lambda_i}{\text{tr}(\Sigma)} = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p} \tag{1.9}$$

Essa medida (1.9) pode ser utilizada como critério para selecionar a quantidade de componentes a serem utilizadas. Vale ressaltar que esse critério é subjetivo, sendo que cabe ao pesquisador definir quantas componentes utilizar, sendo que quanto mais componentes maior a variabilidade explicada.

Ao se trabalhar com um grande número de variáveis é interessante identificar as variáveis mais relevantes, a fim de selecionar um conjunto menor de variáveis a serem analisadas. Utilizando a Análise de Componentes Principais é possível avaliar quais variáveis são mais significativas no banco de dados por meio da interpretação das componentes principais. Usualmente essa interpretação é dada por meio da análise da magnitude dos elementos dos autovetores (cargas vetoriais) $\mathbf{a}_i' = [a_{i1}, \ldots, a_{ip}]$ que representam a correlação entre a i-ésima componente principal e cada uma das variáveis originais, ou seja, a magnitude de a_{i1} representa a importância que $\mathbf{X}_1$ possui pra i-ésima componente $\mathbf{Y}_i$, podendo assim determinar quais variáveis são mais significativas.

Outras informações sobre a Análise de Componentes Principais podem ser encontradas em Johnson e Wichern (2002).

Quando estamos trabalhando com dados superdimensionados, a obtenção das componentes principais utilizando a Análise de Componentes Principais Tradicional se torna árdua por exigir um grande esforço computacional, uma vez que é necessário obter p componentes principais. Além disso os autovetores obtidos se tornam instáveis, sendo que ao retirar ou incluir novas observações ao banco de dados, os elementos dos autovetores, que representam os coeficientes das componentes principais, sofrem grandes modificações. Hastie, Tibshirani e Wainwright (2015).

Com o intuito de resolver esse problema e outras desvantagens da ACP Tradicional, algumas variações da Análise de Componentes Principais foram desenvolvidas e serão apresentadas nas próximas seções.

1.2 Análise de Componentes Principais Esparsas

A não leitura dessa seção não comprometerá o entendimento da aplicação realizada neste trabalho. Uma vez que a utilização dessa técnica não trouxe resultados notavelmente melhores. Entretanto, com o intuito de exemplificar a aplicação dessa técnica em dados reais, é apresentado no final dessa seção um exemplo utilizando os dados *pitprops*, disponíveis no pacote *elasticnet* no software R.

A interpretação das componentes obtidas pela ACP Tradicional se torna mais complexa, em casos que a proporção de variáveis em relação ao número de observações disponíveis é alta, sendo uma das desvantagens dessa técnica quando aplicada a dados superdimensionados. Além disso, um outro problema inerente a técnica é que normalmente

os autovetores estimados não condizem com os reais valores presentes na população.

Com o intuito de solucionar alguns problemas da Análise de Componentes Principais Tradicional, foi desenvolvida uma técnica que trabalha bem quando a quantidade de variáveis é relativamente maior que o tamanho da amostra. Esta técnica ficou conhecida como Análise de Componentes Principais Esparsa (Hastie, Tibshirani e Wainwright (2015)), para lidar com a grande quantidade de variáveis a serem analisadas, Hastie, Tibshirani e Wainwright (2015) impuseram uma restrição de esparsidade nas componentes principais. Essa restrição de esparsidade é dada por meio da utilização das restrições *lasso* (Tibshirani (1996)) ou *elastic net* (Hastie (2000)).

A ACP Esparsa, assim como a ACP Tradicional tem como objetivo reduzir a dimensionalidade dos dados e além disso reduzir o número de variáveis explicitamente utilizadas, a fim de obter uma boa interpretação da estrutura dos dados. São diversas as técnicas em que a interpretação desempenha um papel fundamental para análise dos dados. Como por exemplo em Regressão Linear Múltipla, em que a resposta é predita por uma combinação linear dada pelas variáveis, conhecidas como variáveis explicativas, sendo que um dos objetivos é interpretar o modelo obtido. Usualmente tal interpretação se dá por meio de um critério de seleção de variáveis.

Tibshirani (1996) desenvolveu uma técnica para seleção de variáveis denominada *lasso* (*Least Absolute Shrinkage and Selection Operator*), com o intuito de solucionar alguns problemas na estimação dos coeficientes da Regressão Linear pelo Método dos Mínimos Quadrados. Os estimadores de Mínimos Quadrados não possuem viés, porém alta variabilidade, tornando o poder preditivo do modelo baixo. Além disso, quando o número de variáveis é grande, a interpretação do modelo se torna complexa. A técnica *lasso* consiste em reduzir ou zerar o valor de alguns coeficientes do modelo. Sendo que ao fazer isso, o viés dos estimadores aumenta, porém a variabilidade diminui, aumentando assim o poder preditivo e possibilitando uma melhor interpretação.

O *lasso* funciona como um Método de Mínimos Quadrados Penalizado, em que uma restrição é imposta na norma l_1 dos coeficientes da Regressão. Os estimadores dos coeficientes são obtidos pelo seguinte critério:

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{Y} - \sum_{j=1}^{p} \mathbf{X}_j \beta_j\|^2 + \sum_{j=1}^{p} \lambda |\beta_j|, \tag{1.10}$$

em que λ é não-negativo.

Sendo assim, por meio dessa penalização alguns coeficientes terão seus valores reduzidos e devido à natureza da penalização imposta a norma l_1, se λ for grande o bastante, alguns coeficientes serão nulos. Portanto, o método *lasso* produz simultaneamente um modelo esparso e preciso, tornando-se uma boa técnica para seleção de variáveis. Entretanto, a mesma possui algumas limitações. Uma das mais relevantes, é o fato do número de

variáveis selecionadas pelo método *lasso* depender do tamanho da amostra, sendo possível selecionar no máximo n variáveis, sendo n o tamanho da amostra.

A fim de resolver algumas das limitações do método *lasso*, Hastie (2000) desenvolveram uma técnica chamada *elastic net* que é uma generalização do método *lasso*. Supondo quaisquer λ_1 e λ_2 não negativos, temos que os estimadores para os coeficientes utilizando o método *elastic net*, podem ser obtidos pela fórmula abaixo:

$$\hat{\beta} = (1+\lambda_2)\Big[\arg\min_{\beta} \|\mathbf{Y} - \sum_{j=1}^{p} \mathbf{X}_j\beta_j\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2\Big] \tag{1.11}$$

A penalização *elastic net* é uma combinação convexa das restrições *lasso* e *ridge* (Hoerl e Kennard (1970)) Nota-se que quando $\lambda_2 = 0$, obtemos a penalização *lasso*.

A Análise de Componentes Principais Esparsa foi construída a partir do fato da ACP ser um típico problema de regressão de otimização, em que é imposta uma restrição de esparsidade, seja por *lasso* ou *elastic net*, de tal forma a obter componentes principais esparsas e consequentemente uma boa interpretação. A notação adota por Hastie, Tibshirani e Wainwright (2015) difere de alguns autores, como da adota em Johnson e Wichern (2002), que foi utilizada na seção 1.1. Sendo assim, a seguir serão definidas algumas notações com o intuito de melhorar o entendimento da técnica.

Seja $\mathbf{X}$ uma matriz $n \times p$, sendo que as colunas possuem médias iguais à zero. As componentes principais de $\mathbf{X}$ são obtidas a partir da Decomposição em Valores Singulares (**SVD** – *Singular Value Decomposition*) da matriz $\mathbf{X}$, dada pela seguinte fórmula:

$$\mathbf{X} = \mathbf{UDV}' \tag{1.12}$$

Sendo que as colunas da matriz $\mathbf{U}$, representam os autovetores de $\mathbf{XX}'$, que são conhecidos como vetores singulares à esquerda da matriz $\mathbf{X}$. As colunas da matriz $\mathbf{V}$, representam os autovetores de $\mathbf{X}'\mathbf{X}$ e são chamados de vetores singulares à direita da matriz $\mathbf{X}$. Além disso, os elementos da diagonal de $\mathbf{D}$ $(d_1, d_2, \ldots, d_p)$ são os autovalores de $\mathbf{XX}'$ e $\mathbf{X}'\mathbf{X}$, chamados de valores singulares de $\mathbf{X}$ e são organizados da seguinte maneira $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$.

Os vetores singulares à direita $\mathbf{v}_j$, $j = 1, \ldots, p$, definem as componentes principais de $\mathbf{X}$. Sendo que a partir do vetor unitário $\mathbf{v}_1$ é possível obter a combinação linear $Y_1 = \mathbf{X}\mathbf{v}_1$, que possui a maior variabilidade dentre todas as possíveis combinações lineares. Portanto, Y_1 é chamada de primeira componente principal, e $\mathbf{v}_1$ é conhecida como carga vetorial ou direção. Similarmente, $Y_2 = \mathbf{X}\mathbf{v}_2$, representa a segunda componente principal com maior variabilidade, sujeita a não ser correlacionada com Y_1. Seguindo esta lógica é possível obter k componentes principais de $\mathbf{X}$.

A ACP Esparsa impõe esparsidade na forma em que a variância é maximizada na Análise de Componentes Principais, por meio da imposição de uma restrição na norma l_1:

$$\underset{\|\mathbf{v}_1\|_2=\|\mathbf{u}_1\|_2=1}{\arg\max} \{\mathbf{u}_1'\mathbf{X}\mathbf{v}_1\} \text{ sujeito a } \|\mathbf{v}_1\|_1 \leq t \tag{1.13}$$

em que t é um parâmetro de ajuste. Temos que para $t \geq \sqrt{p}$ obtemos ACP Tradicional, $t < 1$ não temos solução, e para $t = 1$ teremos exatamente uma carga não-nula em cada componente.

Devido ao fato da equação (1.13) ser um problema bi-convexo no par $(\mathbf{u}_1, \mathbf{v}_1)$, podemos aplicar o seguinte algoritmo com o intuito de resolvê-lo.

1. Inicia-se $\mathbf{v}_1 \in \mathbb{R}^p$ com $\|\mathbf{v_1}\|_2 = 1$
2. Este processo é repetido até que as mudanças em $\mathbf{u}_1$ e $\mathbf{v}_1$ sejam suficientemente pequenas.
   a) Atualiza-se $\mathbf{u}_1 \in \mathbb{R}^n$, $\mathbf{u}_1 \leftarrow \frac{\mathbf{X}\mathbf{v}_1}{\|\mathbf{X}\mathbf{v}_1\|_2}$
   b) Atualiza-se $\mathbf{v}_1 \in \mathbb{R}^p$, $\mathbf{v}_1 \leftarrow \mathbf{v}_1(\lambda, \mathbf{u}) = \frac{S_\lambda(\mathbf{X}'\mathbf{u})}{\|S_\lambda(\mathbf{X}'\mathbf{u})\|_2}$

   em que $S_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+$,

   Escolhe-se $\lambda = 0$ se $\|\mathbf{X}'\mathbf{u}\|_1 \leq t$ e $\lambda > 0$ se $\|\mathbf{v_1}(\lambda, \mathbf{u})\|_1 = t$.

Por meio do algoritmo acima, obtemos $(\mathbf{u}_1, \mathbf{v}_1, d_1)$. Calcula-se então o resíduo $\mathbf{X}' = \mathbf{X} - d_1\mathbf{u}_1\mathbf{v}_1'$ e aplicando o algoritmo novamente, obtemos a segunda solução $(\mathbf{u}_2, \mathbf{v}_2, d_2)$, pode-se repetir este processo até que se obtenha k soluções $(\mathbf{u}_1, \mathbf{v}_1, d_1), \ldots, (\mathbf{u}_k, \mathbf{v}_k, d_k)$. A partir dessas soluções pode-se determinar as componentes principais esparsas, $Y_j = \mathbf{u}_j d_j = \mathbf{X}\mathbf{v}_j$, $j = 1, \ldots, k$. Vale ressaltar que a ACP esparsa não garante ortogonalidade entre as componentes principais e nem para cargas vetoriais esparsas $(\mathbf{v}_1, \ldots, \mathbf{v}_k)$. Entretanto, na prática as soluções tendem a serem aproximadamente ortogonais. Algumas técnicas como a $SCoTLASS$ forçam que as componentes principais sejam ortogonais, porém ao se fazer isso, obtém-se como resultado componentes menos esparsas.

A equação (1.13) pode ser modificada para que os vetores $\mathbf{u_j}$ sejam ortogonais, sem que seja colocada uma restrição nos vetores $\mathbf{v}_j$, sendo possível assim garantir a ortogonalidade e a esparsidade.

$$\underset{\mathbf{u}_k, \mathbf{v}_k}{\arg\max} \{\mathbf{u}_k'\mathbf{X}\mathbf{v}_k\} \text{ sujeito a } \|\mathbf{v}_k\|_2 \leq 1, \quad \|\mathbf{v}_k\|_1 \leq c, \tag{1.14}$$

$$\|\mathbf{u_k}\|_2 \leq 1 \text{ com } \mathbf{u}_k'\mathbf{u}_j = 0, \quad \forall j = 1, \ldots, k-1$$

A solução para $\mathbf{u}_k$, com $\mathbf{v}_k$ fixo é dada por:

$$\mathbf{u}_k = \frac{\mathbf{P}^{\perp}{}_{k-1}\mathbf{X}\mathbf{v}_k}{\|\mathbf{P}^{\perp}{}_{k-1}\mathbf{X}\mathbf{v}_k\|_2} \tag{1.15}$$

em que,

$\mathbf{P}^{\perp}{}_{k-1} = \mathbf{I} - \sum_{i=1}^{k-1} \mathbf{u}_i\mathbf{u}_i'$. Sendo que para obter as componentes principais ortogonais e esparsas, basta substituir $\mathbf{u}_1$ no algoritmo apresentado anteriormente por $\mathbf{u}_k$

Paralelamente, o autovetor $\mathbf{v}_k$, pode ser visto como o autovetor $\mathbf{a}_k$ da Análise de Componentes Principais Tradicional, porém $\mathbf{v}_k$ é esparso. Como visto, na ACP Esparsa são obtidas k componentes principais, sendo que cada componente pode possuir um número diferente de coeficientes (cargas) não nulas.

Para exemplificar a utilização da Análise de Componentes Principais Esparsa em dados reais, utilizou-se os dados *pitprops*, presentes no pacote *elasticnet* no Software R. A base de dados *pitprops* contém a matriz de correlação de 13 medidas referentes à 180 cortes de madeira utilizadas em minas (Tabela 1). Jeffers (1967), utilizando esta base de dados tentou interpretar as seis primeiras componentes por meio da aplicação da Análise de Componentes Principais Tradicional e deparou-se com a dificuldade em interpretar as componentes quando possuem cargas não-nulas. Sendo assim, Hastie, Tibshirani e Wainwright (2015) aplicaram a Análise de Componentes Principais Esparsa, com o intuito de mostrar que a mesma facilita a interpretação das componentes obtidas quando impõe-se uma penalização nas coeficientes das componentes, tornando-as esparsas.

Tabela 1 – Pitprops

Medida	Significado
`topdiam`	Diâmetro superior em polegadas
`length`	Comprimento em polegadas
`moist`	Teor de umidade, % do peso seco
`testsg`	Gravidade específica durante o teste
`ovensg`	Gravidade específica seco ao forno
`ringtop`	Número de anéis anuais na parte superior
`ringbut`	Número de anéis anuais na parte inferior
`bowmax`	Arco máximo em polegadas
`bowdist`	Distância de cima até arco máximo em polegadas
`whorls`	Número de espirais de nós
`clear`	Comprimento de suporte da parte superior em polegadas
`knots`	Número médio de nós por espiral
`diaknot`	Diâmetro médio dos nós em polegadas

Assim como Jeffers (1967), Hastie, Tibshirani e Wainwright (2015) consideraram as primeiras seis componentes principais e a penalização adotada foi escolhida de tal forma

que as componentes explicassem aproximadamente a mesma proporção da variabilidade que as obtidas pelas ACP Tradicional.

Tabela 2 – Pitprops - ACP

Variável	CP1	CP2	CP3	CP4	CP5	CP6
topdiam	0,4038	0,2178	0,2073	0,0912	0,0826	0,1198
length	0,4055	0,1861	0,2350	0,1027	0,1128	0,1629
moist	0,1244	0,5406	-0,1415	-0,0784	-0,3498	-0,2759
testsg	0,1732	0,4556	-0,3524	-0,0548	-0,3557	-0,0540
ovensg	0,0572	-0,1701	-0,4812	-0,0491	-0,1761	0,6255
ringtop	0,2844	-0,0142	-0,4752	0,0634	0,3158	0,0523
ringbut	0,3998	-0,1896	-0,2531	0,0649	0,2151	0,0026
bowmax	0,2935	-0,1891	0,2430	-0,2855	-0,1853	-0,0551
bowdist	0,3566	0,0171	0,2076	-0,0967	0,1061	0,0342
whorls	0,3789	-0,2484	0,1188	0,2050	-0,1564	-0,1731
clear	-0,0111	0,2053	0,0704	-0,8036	0,3429	0,1753
knots	-0,1151	0,3432	-0,0919	0,3008	0,6004	-0,1698
diaknot	-0,1125	0,3085	0,3261	0,3034	-0,0799	0,6263
variabilidade (%)	32,45	18,29	14,45	8,53	7,00	6,27
Cargas não-nulas	13	13	13	13	13	13

Tabela 3 – Pitprops - ACP Esparsa

Variável	CP1	CP2	CP3	CP4	CP5	CP6
topdiam	-0,4774	0	0	0	0	0
length	-0,4759	0	0	0	0	0
moist	0	0,7847	0	0	0	0
testsg	0	0,6194	0	0	0	0
ovensg	0,1766	0	0,6407	0	0	0
ringtop	0	0	0,5890	0	0	0
ringbut	-0,2505	0	0,4923	0	0	0
bowmax	-0,3440	-0,0210	0	0	0	0
bowdist	-0,4164	0	0	0	0	0
whorls	-0,4000	0	0	0	0	0
clear	0	0	0	-1	0	0
knots	0	0,0133	0	0	-1	0
diaknot	0	0	-0,0156	0	0	1
variabilidade (%)	28,03	13,96	13,30	7,44	6,80	6,22
Cargas não-nulas	7	4	4	1	1	1

As tabelas 2 e 3 contêm os resultados obtidos com a aplicação da ACP Tradicional e ACP Esparsa utilizando os dados *pitprops*. Sendo possível perceber que mesmo explicando menos da variabilidade dos dados com as seis componentes principais (75,77%), a ACP Esparsa obteve componentes com cargas nulas, facilitando a interpretação das mesmas.

1.3 Análise de Componentes Principais Bidimensional

A Análise de Componentes Principais pode ser aplicada em diversas áreas, tendo como objetivo explicar a estrutura dos dados. Ao se trabalhar com reconhecimento de padrões em imagens 2D, a Análise de Componentes Principais necessita de algumas modificações. Sendo assim, Dwivedi (2008) apresentou uma extensão da ACP Tradicional para representação de imagens 2D, conhecida como Análise de Componentes Principais Bidimensional (2DPCA).

Uma das abordagens da Análise de Componentes Principais Bidimensional consiste em dividir uma imagem em blocos de dimensão $b \times b$ pixels, em que cada bloco é transformado em um vetor. Após esse processo, a matriz de covariâncias desse conjunto de vetores é obtida e aplica-se então a ACP Tradicional nesses dados. Feito isso é possível recuperar a imagem por meio das componentes obtidas, possuindo preferencialmente uma baixa dimensão. (Diagrama 1)

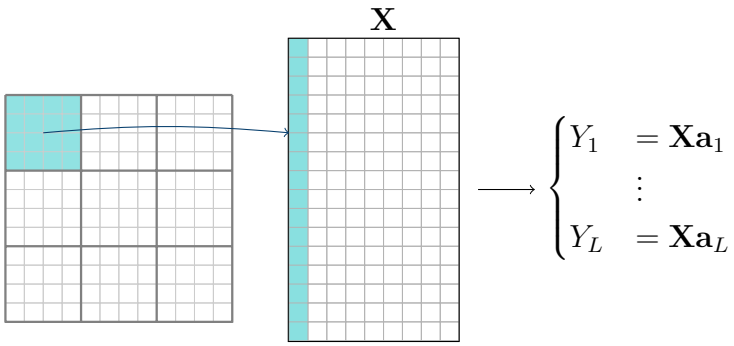


Figura 1 – Estrutura da Análise de Componentes Principais Bidimensional

Como exemplificação, suponha que uma imagem tenha dimensão 512×512 pixels, dividindo a mesma em blocos de 8×8 pixels e transformando-os em vetores, é possível representar a imagem por meio de 4096 vetores em $\mathbb{R}^{64}$.

Seja $\mathbf{A}$ uma imagem de dimensão $m \times n$, em que a mesma é dividida em L blocos de dimensão $b \times b$. Feito isso, cada bloco é transformado em um vetor $\mathbf{x}_i$ presente em $\mathbb{R}^{b^2}$.

$$\mathbf{A} = \begin{bmatrix} z_{11} & z_{12} & z_{13} & \cdots & z_{1n} \\ z_{21} & z_{22} & z_{23} & \cdots & z_{2n} \\ z_{31} & z_{32} & z_{33} & \cdots & z_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{m1} & z_{m2} & z_{m3} & \cdots & z_{mn} \end{bmatrix}_{m \times n}$$

$$\mathbf{A} = \left[\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1b} & b_{11} & \cdots & b_{1b} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{b1} & \cdots & a_{bb} & b_{b1} & \cdots & b_{bb} \\ \hline c_{11} & \cdots & c_{1b} & d_{11} & \cdots & d_{1b} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c_{b1} & \cdots & c_{bb} & d_{b1} & \cdots & d_{bb} \end{array}\right]_{m \times n} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L]$$

Por meio dos vetores obtidos é possível representar a imagem por $\mathbf{A} = \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$. Após esse processo, a matriz de covariâncias $\Sigma_{\mathbf{x}}$ é obtida utilizando a decomposição $\Sigma_{\mathbf{x}} = \Gamma_{\mathbf{x}} \Lambda_{\mathbf{x}} \Gamma_{\mathbf{x}}'$. Em que:

$$\Gamma_{\mathbf{x}} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L] = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{L1} \\ \vdots & \ddots & \vdots \\ \gamma_{1L} & \cdots & \gamma_{LL} \end{bmatrix}, \quad \Lambda_{\mathbf{x}} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_L \end{bmatrix},$$

$\lambda_1, \ldots, \lambda_L$ são os autovalores de $\Sigma_{\mathbf{x}}$, sendo $\lambda_1 \geq \cdots \geq \lambda_L \geq 0$ e $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L$ representam os autovetores associados aos autovalores de $\Sigma_{\mathbf{x}}$.

A partir dos autovetores $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L$ é possível obter as componentes principais ($\mathbf{Y} = \mathbf{X}\Gamma_{\mathbf{x}}$) e reconstruir a imagem utilizando essas componentes. Sendo assim, a imagem aproximada $\tilde{\mathbf{A}}$, pode ser obtida por meio da seguinte fórmula:

$$\tilde{\mathbf{A}} = [\mathbf{y}_1 \Gamma_{\mathbf{x}}', \ldots, \mathbf{y}_k \Gamma_{\mathbf{x}}'] \tag{1.16}$$

Essa técnica possui algumas limitações ao se trabalhar com imagens de alta dimensão, usualmente imagens com dimensão maior que 100×100 pixels, pois os vetores gerados pelos blocos são superdimensionados, tornando a estimação da matriz de covariâncias bastante complexa e muitas vezes não estimável.

Existem diversas variações da técnica 2DPCA, tais como: **2DPCA alternativo**, **(2D)^{2}PCA** e **DiaPCA**, que são aplicadas para casos específicos e são explicadas em Yang et al. (2004).

Vimos que em casos que a dimensão da imagem cresce consideravelmente, a ACP Tradicional possui limitações, fazendo com que muitas vezes não seja possível aplicar a mesma diretamente, tornando a Análise de Componentes Principais Esparsa e a Análise de Componentes Principais Bidimensional boas opções para análise de imagens. No caso, a ACP Esparsa é bastante utilizada quando as imagens possuem grande quantidade de zeros, caracterizadas como esparsas. Neste caso, a ACP Esparsa além de reduzir a dimensionalidade das imagens, facilita também a interpretação das componentes obtidas.

Enquanto, a 2DPCA consiste em aplicar a Análise de Componentes Principais, seja ela Tradicional ou Esparsa, em uma matriz de dados não superdimensionada, por meio da divisão das imagens em blocos, com o intuito em reconstruir a imagem em uma dimensão mais baixa.

2 Análise da Digestão por Cintilografia

Neste Capítulo serão apresentados alguns resultados obtidos por meio da aplicação da Análise de Componentes Principais para identificação de regiões de interesse, limpeza e análise de imagens por cintilografia do estômago. Primeiramente, serão apresentados alguns tópicos em relação as imagens. Em seguida, serão apresentados os resultados obtidos.

2.1 Imagens no padrão DICOM

Com a crescente utilização de computadores na área médica, como na coleta de exames, visualização de imagens, compartilhamento de resultados, a American College of Radiology e a National Electrical Manufacturers Association, desenvolveram um padrão com o intuito de armazenar e compartilhar informações médicas em um formato único, geradas a partir de equipamentos médicos, permitindo ter um diagnóstico mais detalhado e a distância. Esse padrão foi intitulado como padrão DICOM (Digital Imaging and Communications in Medicine).

O padrão DICOM visa a troca de informações digitais entre equipamentos de imagens médicas e outros sistemas, sendo desenvolvido com ênfase no diagnóstico de imagens praticadas em radiologia, cardiologia, patologia, odontologia, oftalmologia, dentre outras áreas.

Thornton (2011) desenvolveu um pacote na linguagem de programação estatística R (*oro.dicom*), com o intuito de facilitar a manipulação de imagens médicas no padrão DICOM, principalmente na importação/exportação e na visualização de imagens.

2.2 Conjunto de dados

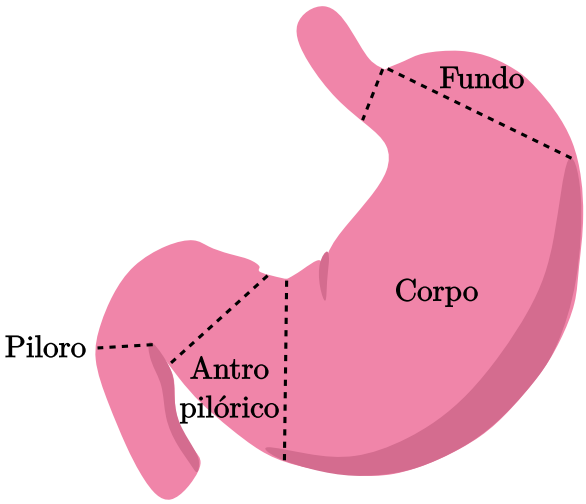
O conjunto de dados utilizado para aplicação das técnicas abordadas neste trabalho foi coletado por um grupo de pesquisadores da seção de Gastroenterologia da Universidade Texas Tech em El Paso, liderados pelo Dr. Richard W. McCallum. Os dados foram pré-processados pelo Dr. Ricardo von Borries, do departamento de Engenharia Elétrica e Computacional da Universidade do Texas em El Paso. O banco de dados, em formato `csv` é composto por imagens cintilográficas do estômago de 6 (seis) indivíduos, cada imagem possui dimensão 64×64 pixels e foram coletadas em 8 (oito) intervalos de tempo: 15, 30, 45, 60, 90, 120, 180 e 240 minutos após a ingestão de uma refeição. Além disso, em cada intervalo de tempo foram coletados 512 frames (sequência de imagens), as imagens foram coletadas originalmente no padrão DICOM, descrito na seção 2.1. A princípio pensou-se

em utilizar o pacote *oro.dicom* (Thornton (2011)) para realizar a leitura das imagens, porém devido a existência de algumas dúvidas acerca dos resultados obtidos com o pacote, as imagens foram convertidas para o formato CSV (*Comma-separated values*) para que fosse realizada a leitura pelo R.

A análise desses dados é de suma importância para que se possa entender os mecanismos responsáveis pelos padrões da motilidade gástrica. Sendo possível diagnosticar distúrbios, tais como a Gastroparesia, uma condição que afeta os músculos do estômago e impede o esvaziamento adequado deste órgão, causando dores e outras complicações.

Anatomicamente, o estômago é composto basicamente por quatro partes sendo elas: o fundo, corpo, antro pilórico e piloro (Figura 2). A função motora do estômago consiste na absorção de alimentos e na eliminação de resíduos. Um dos principais estudos realizados para avaliar a motilidade gástrica, consiste em analisar o esvaziamento gástrico, que tem como intuito identificar, caracterizar e estabelecer possíveis correlações fisiopatológicas. Além disso, por meio do estudo do esvaziamento gástrico é possível obter informações sobre a velocidade do processo digestivo, sendo uma importante ferramenta para diagnósticos de distúrbios gastrointestinais.

Figura 2 – Estômago



Fonte: Imagem original por Lina Wolf, Wikimedia Commons. Modificada pela autora.

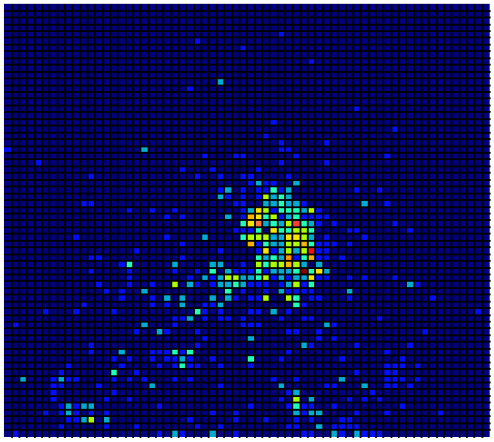
Atualmente existem diversas técnicas que avaliam o esvaziamento gástrico, buscando definir a taxa de eliminação de um determinado conteúdo gástrico. Dentre elas, a mais utilizada é a cintilografia, sendo uma técnica não invasiva, que consiste na obtenção de imagens por meio de varreduras, realizadas após a ingestão de uma refeição radiomarcada, acompanhando a digestão dessa refeição nas primeiras quatro horas após a ingestão da mesma. Essa refeição teste deve possuir um conteúdo calórico suficiente (maior que 200 calorias) e consistência sólida, de tal forma que induza o aumento de contrações do estômago. A refeição-teste normalmente utilizada é composta por um sanduíche de ovos

com coloide de enxofre e tecnécio-99m (Tc-99m).

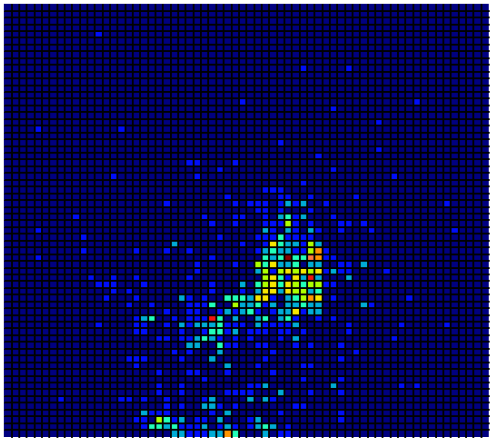
As imagens foram obtidas por meio da radiação emitida pelo radiofármaco Tc-99m, colocado no alimento para servir de contraste. Esse radiofármaco interage com o aparelho que detecta a radiação, produzindo emissão de luz, sendo possível identificar a posição em que está sendo emitida e a sua intensidade, os pixels da imagem são determinados por essa intensidade, afetando a coloração da imagem.

Figura 3 – Comparação da primeira e da última imagens das sequências obtidas em 15 e 60 minutos após a ingestão

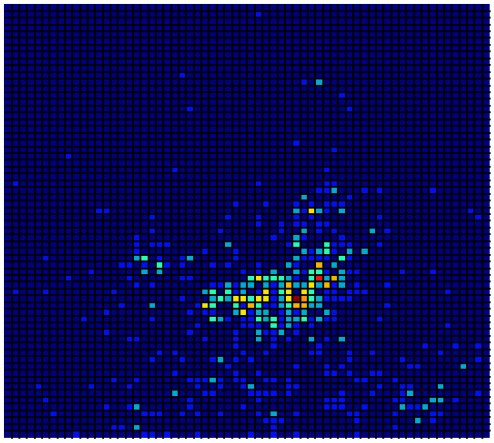
(a) Frame 1 em 15 minutos após ingestão



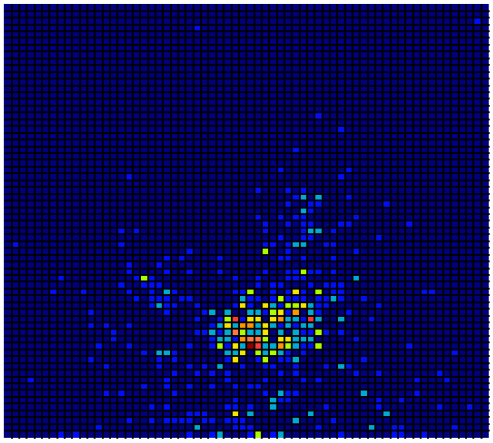
(b) Frame 512 em 15 minutos após ingestão



(c) Frame 1 em 60 minutos após ingestão



(d) Frame 512 em 60 minutos após ingestão

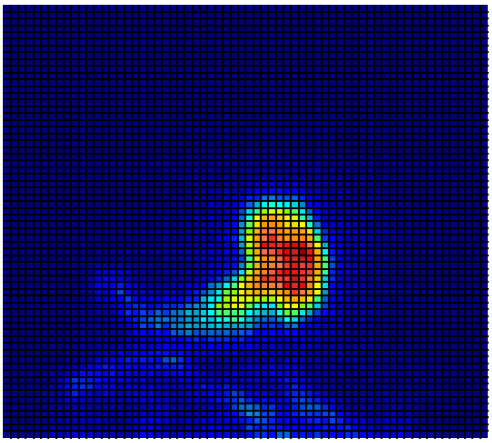


Para que se pudesse ter uma ideia das imagens utilizadas neste trabalho, foram expostas na Figura 3 a primeira (Frame 1) e a última (Frame 512) imagens das sequências obtidas nos tempos 15 e 60 minutos após a ingestão da refeição radiomarcada. Nota-se

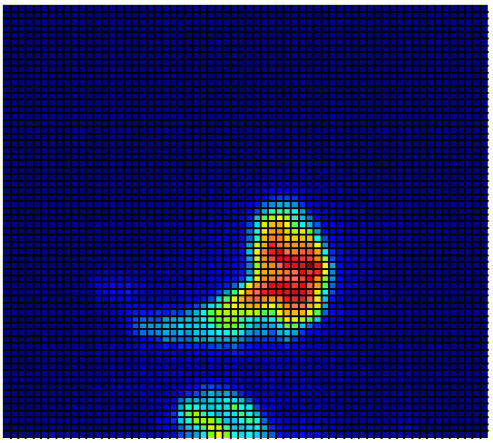
uma pequena diferença ao se comparar as duas imagens em cada tempo. Tendo em vista isso, podemos pensar em trabalhar com uma quantidade menor de imagens, subdividindo a sequencia de imagens (512 frames) em dois grupos, sendo o primeiro composto pelas 256 primeiras imagens e o segundo pelas 256 restantes e calculando a média para cada grupo foi possível obter duas imagens representativas para cada tempo (Figura 4).

Figura 4 – Imagens médias para os grupos de imagens nos tempos 15 e 60 minutos

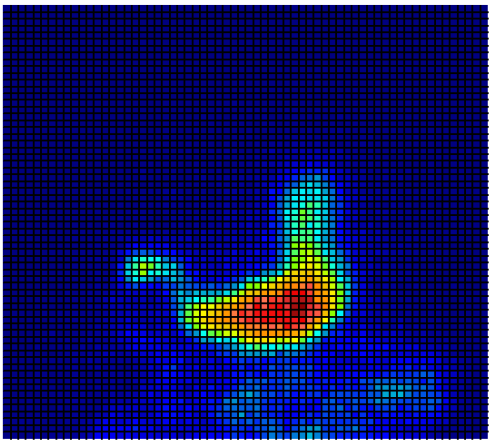
(a) Grupo 1 em 15 minutos



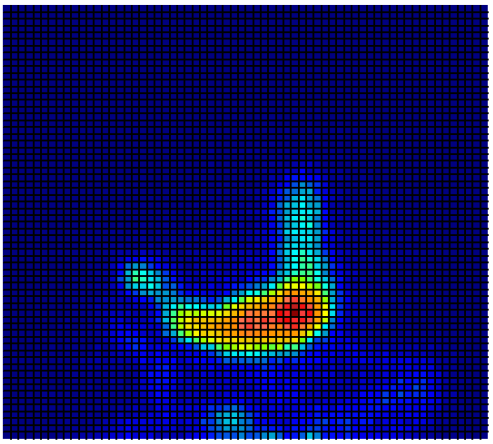
(b) Grupo 2 em 15 minutos



(c) Grupo 1 em 60 minutos



(d) Grupo 2 em 60 minutos



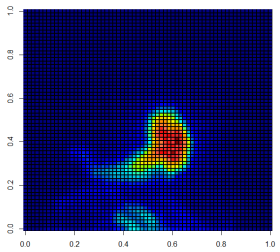
Percebe-se pela Figura 4 que, ao se dividir a sequência de imagens em dois grupos e calcular a imagem média em cada grupo, é possível ter uma visualização mais fácil da informação contida nas imagens. Dado que, os frames obtidos originalmente acabam dificultando a visualização da região que representaria o estômago (Figura 3).

Sendo assim, para visualização das imagens em cada ponto no tempo, optou-se em utilizar as imagens médias obtidas, pois facilitaram a visualização da informação contida nas mesmas.

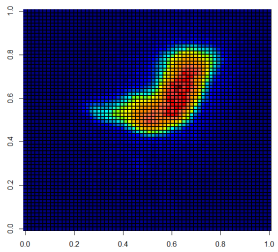
Figura 5 – Comparação do comportamento digestivo entre os indivíduos 2 e 4

Indivíduo 2 **Indivíduo 4**

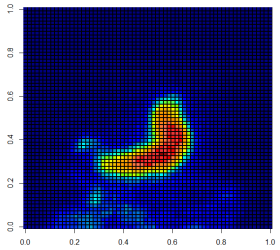
(a) 15 min após ingestão



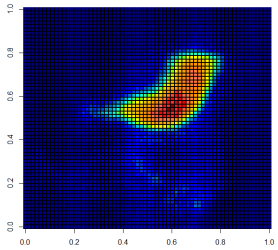
(b) 15 min após ingestão



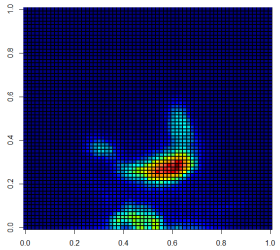
(c) 45 min após a ingestão



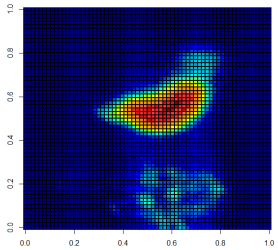
(d) 45 min após a ingestão



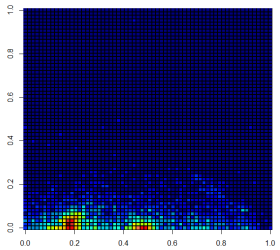
(e) 90 min após a ingestão



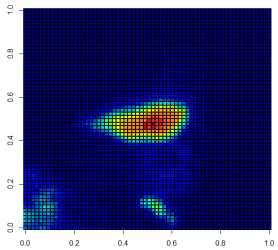
(f) 90 min após a ingestão



(g) 240 min após a ingestão



(h) 240 min após a ingestão



Ao se avaliar a motilidade gástrica, é intuitivo pensar que o comportamento digestivo entre indivíduos pode variar, sendo interessante determinar se o padrão digestivo dos mesmos se encontra normal ou possui alguma anormalidade, podendo ser justificado por uma má funcionalidade do estômago, interferindo diretamente a motilidade gástrica. Portando, a análise da motilidade auxilia na detecção de anomalias na função gástrica do estômago, servindo de ferramenta para diagnósticos de possíveis distúrbios gastrointestinais. Além disso, por meio da análise do comportamento digestivo dos indivíduos é possível identificar e agrupar indivíduos com padrões semelhantes.

Sendo assim, com o intuito de verificar semelhanças entre os padrões digestivos dos indivíduos, foram expostas na Figura 5 as imagens médias em quatro pontos no tempo (15, 45, 90 e 240 minutos após a ingestão da refeição radiomarcada) para dois indivíduos, sendo que os gráficos à esquerda correspondem ao indivíduo 2 e os à direita correspondem ao indivíduo 4.

Nota-se, que no início (15 minutos) os dois indivíduos possuíam um padrão semelhante, o alimento encontrava-se concentrado, porém à medida que o tempo foi passando o comportamento digestivo dos dois indivíduos começou a diferir, observando que aos 90 e 240 minutos após a ingestão da refeição, o indivíduo 2 tinha praticamente digerido completamente o alimento. Entretanto, o indivíduo 4 não conseguiu digerir completamente o alimento. Essa informação nos leva a pensar que o indivíduo 4 possui um esvaziamento gástrico mais lento quando comparado ao indivíduo 2.

Ainda pela Figura 5, é possível perceber que as imagens não representam somente o estômago, que é a região de interesse do estudo, mas também outras regiões próximas ao estômago, como o intestino. Sendo assim, como o foco do estudo consiste em avaliar a motilidade do estômago é necessário identificar primeiramente a região de interesse (Figura 2), para assim avaliar a motilidade gástrica dos indivíduos.

Logo, um dos objetivos deste trabalho consiste em identificar a região de interesse de forma automática para diferentes intervalos de tempo, considerando o fato que a posição do indivíduo em diferentes intervalos pode estar relativamente deslocada. Após a identificação da região de interesse, temos como objetivo ainda analisar o comportamento espaço-temporal da digestão dos indivíduos, identificando os marcadores do movimento do alimento no estômago para análise futura.

2.3 Estudo I – Limpeza das imagens

Na seção 1.3 vimos que ao se trabalhar com imagens, uma técnica bastante utilizada é a Análise de Componentes Principais Bidimensional (2DPCA), sendo uma extensão da Análise de Componentes Principais Tradicional (ACP Tradicional) para reconhecimento de padrões em imagens 2D. A 2DPCA consiste em dividir a imagem em blocos de dimensão $b \times b$, transformá-los em vetores, obter a matriz de covariâncias desse conjunto de vetores e aplicar a ACP Tradicional, de tal forma que seja possível reconstruir a imagem com uma baixa dimensão. Vale ressaltar que, poderíamos utilizar a Análise de Componentes Principais Esparsa ao em vez da Tradicional, pois as imagens são caracterizadas como esparsas. Entretanto, não observou-se resultados notavelmente melhores. Sendo assim, optou-se em trabalhar com a Análise de Componentes Principais Tradicional.

Como mencionado anteriormente, as imagens obtidas por cintilografia acabam captando informação sobre áreas próximas ao estômago. Como o objetivo deste trabalho consiste em avaliar o comportamento digestivo de determinados indivíduos é necessário desconsiderar a informação dada pelas regiões que não correspondem ao estômago, realizando uma limpeza nas imagens a serem analisadas, facilitando a identificação da região de interesse.

Para realização da limpeza das imagens, subdividiu-se as 512 imagens de cada ponto no tempo em dois grupos, sendo o primeiro composto pelas primeiras 256 imagens e o segundo com as restantes. Feito isso, foi calculada a mediana das imagens em cada grupo, obtendo assim duas imagens representativas para cada ponto no tempo, observou-se que com a utilização da mediana foi possível realizar a limpeza das imagens, sem que fosse necessário utilizar a Análise de Componentes Principais.

Na Figura 6, temos os resultados obtidos com a aplicação da mediana nas imagens de um determinado indivíduo para cada ponto no tempo, foi possível observar que ao final das quatro horas de acompanhamento, o indivíduo em questão havia digerido completamente o alimento.

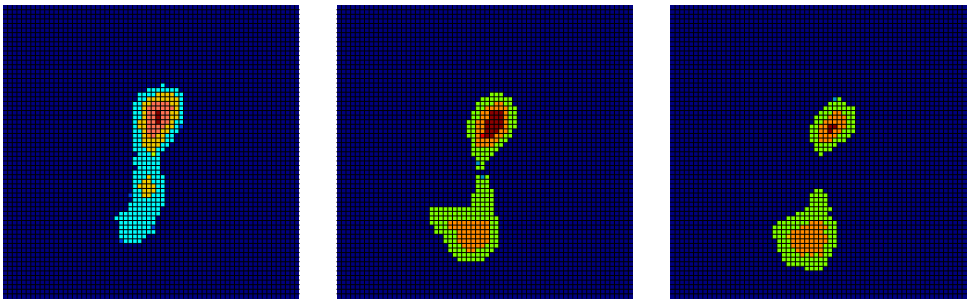
2.4 Estudo II – Identificação da região de interesse

Normalmente, a região de interesse é determinada de forma não automática, sendo necessário desenhar a mesma em cada imagem gerada. Este procedimento é necessário, pois as imagens obtidas por meio da cintilografia acabam captando informação sobre áreas próximas ao estômago, além do fato da posição do indivíduo poder estar levemente deslocada em diferentes intervalos de tempo.

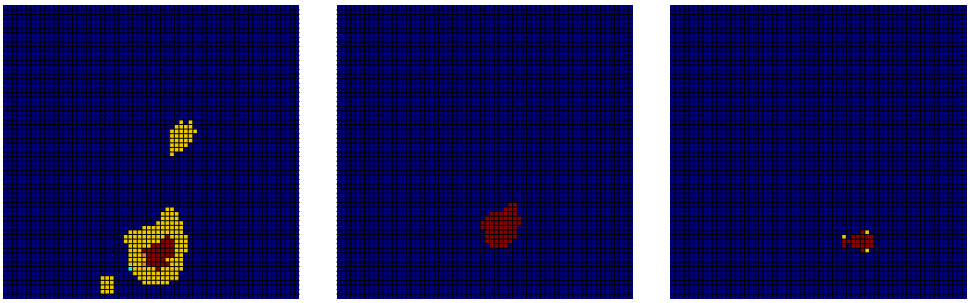
Pela Figura 6 é possível perceber que com a utilização da mediana além de ter sido possível limpar as imagens, foi possível identificar a região de interesse, uma vez que

Figura 6 – Limpeza das imagens para todos os pontos no tempo indivíduo 1

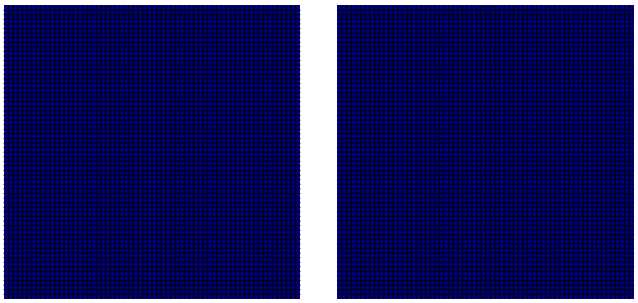
(a) 15 minutos (b) 30 minutos (c) 45 minutos



(d) 60 minutos (e) 90 minutos (f) 120 minutos



(g) 180 minutos (h) 240 minutos



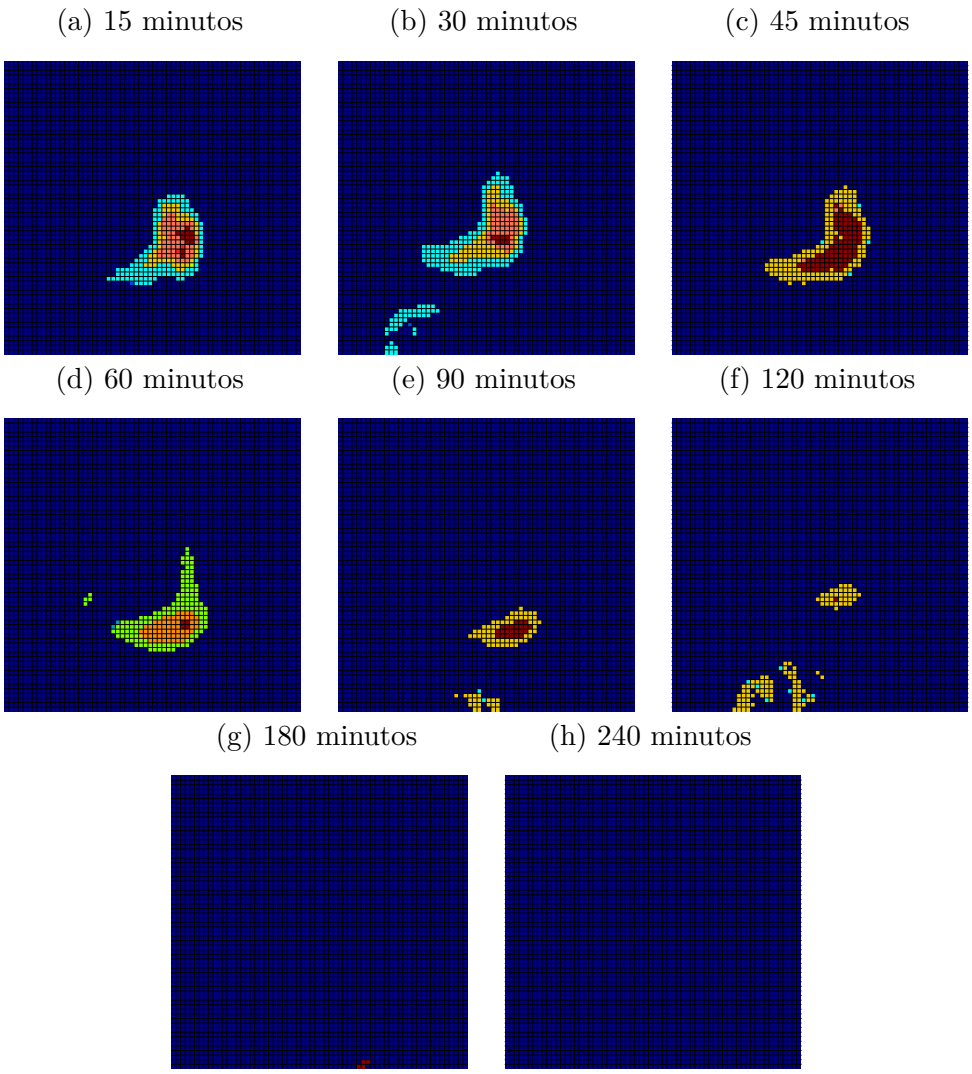
com a mediana a informação presente nas imagens das regiões próximas ao estômago foi retirada, restando somente a região de interesse, representada pelo estômago.

Com o intuito de expor os resultados obtidos com a identificação da região de interesse por meio da utilização da mediana, foram colocadas nas Figuras 7 e 8 as regiões de interesse obtidas em cada ponto no tempo para dois indivíduos.

Uma vez identificada a região de interesse para cada ponto no tempo e para todos os indivíduos, foi avaliada a velocidade da digestão de cada indivíduo por meio da utilização da Análise de Componentes Principais, com esta informação é possível identificar possíveis distúrbios gastrointestinais.

Sendo que, esta informação pode ser obtida por meio da análise do número de componentes principais necessárias para explicar um percentual (alto) da variabilidade dos dados em cada ponto no tempo. Sendo intuitivo pensar que a medida que o tempo passa, o alimento é digerido, com isso a informação contida em cada imagem diminui ao

Figura 7 – Identificação da região de interesse para o indivíduo 2



longo do tempo e consequentemente o número de componentes.

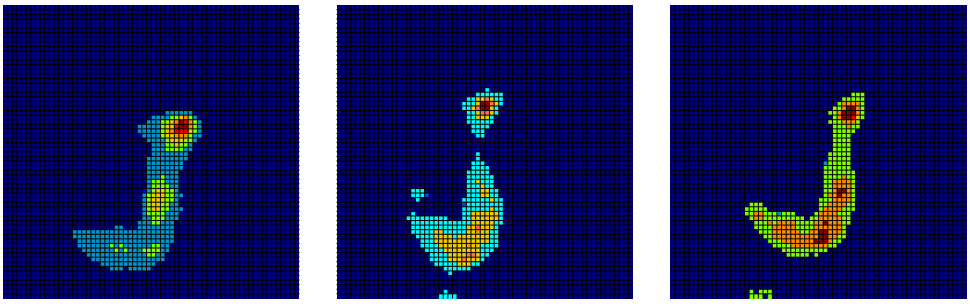
2.5 Estudo III – Velocidade da Digestão

A determinação da velocidade do processo digestivo se deu por meio da Análise de Componentes Principais Bidimensional, em que foram obtidos gráficos contendo informação sobre o número de componentes necessárias para explicar 95% da variabilidade contida nas imagens medianas em cada ponto no tempo e para cada indivíduo. É intuitivo pensar que à medida que o tempo passa, o número de componentes deverá reduzir, pois o alimento está sendo digerido, consequentemente há menos informação nas imagens, influenciando o número de componentes obtidas para explicar 95% da variabilidade total. Por meio desses gráficos foi possível ter uma ideia do tempo que o indivíduo levou para digerir o alimento.

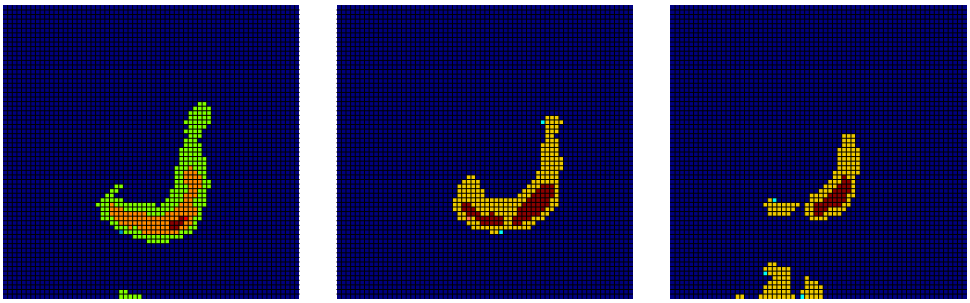
Com essa informação é possível identificar se o processo digestivo se encontra lento, normal ou rápido. Sendo que distúrbios gastrointestinais podem ser ocasionados quando o

Figura 8 – Identificação da região de interesse para o indivíduo 6

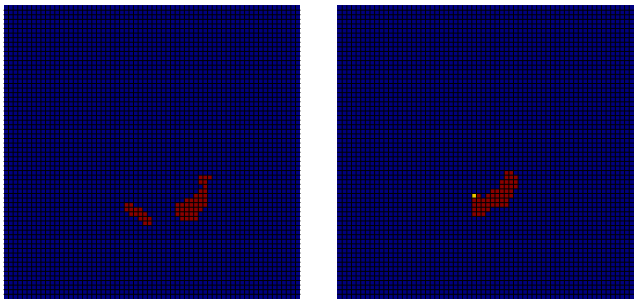
(a) 15 minutos (b) 30 minutos (c) 45 minutos



(d) 60 minutos (e) 90 minutos (f) 120 minutos



(g) 180 minutos (h) 240 minutos



esvaziamento gástrico se encontra anormal (lento ou rápido). Malmud et al. (1982).

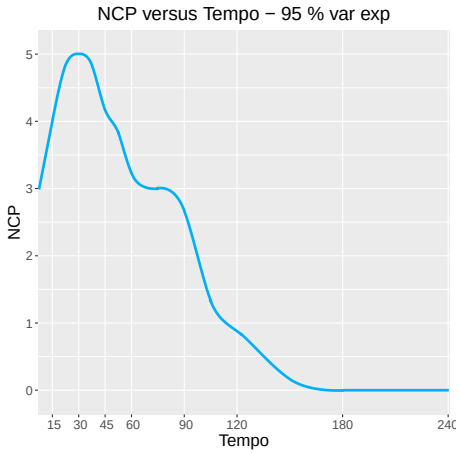
O esvaziamento gástrico em taxa lenta pode ser ocasionado por uma obstrução mecânica ou funcional. A obstrução mecânica resulta do estreitamento ou entupimento do lúmen, reduzindo o fluxo de saída do alimento para o intestino delgado. Enquanto, a obstrução funcional resulta de uma anormalidade da motilidade gástrica associada a dificuldade na digestão de alimentos sólidos ou na inabilidade em eliminar alimentos do estômago.

As principais causas que podem resultar em um esvaziamento gástrico em taxa lenta são: Câncer de estômago, Gastroparesia, Hipotireoidismo e Úlcera gástrica.

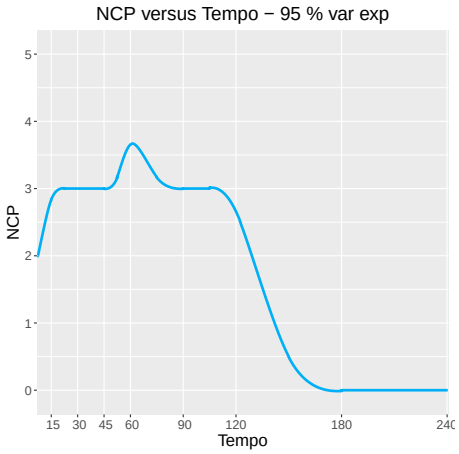
O esvaziamento gástrico em taxa rápida ocorre menos frequentemente e usualmente é iatrogênico, ou seja, resulta de complicações causadas por um procedimento médico. Entretanto, Hipotireoidismo, Úlcera Duodenal podem induzir o esvaziamento em taxa rápida.

Figura 9 – Velocidade da digestão

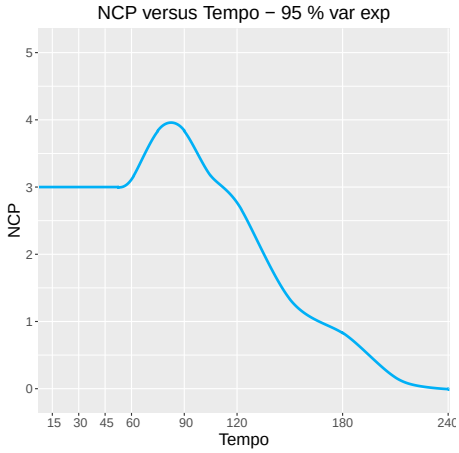
(a) Indivíduo 1



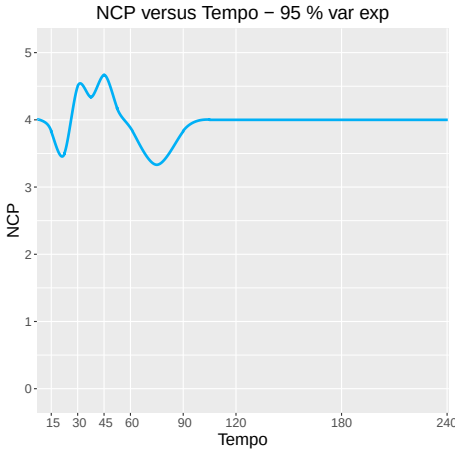
(b) Indivíduo 2



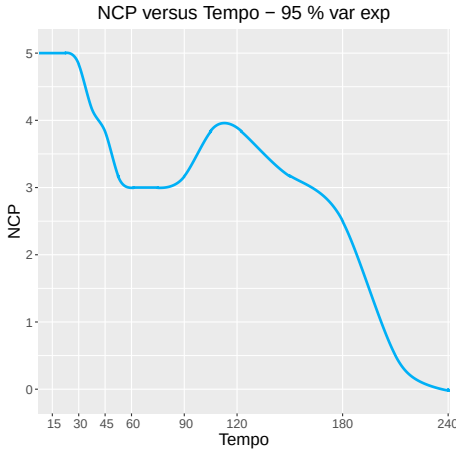
(c) Indivíduo 3



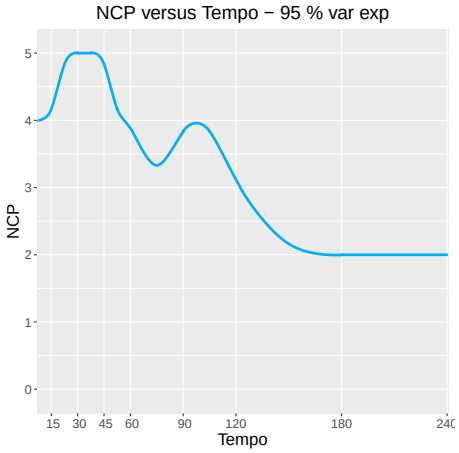
(d) Indivíduo 4



(e) Indivíduo 5



(f) Indivíduo 6



Sendo assim, foram gerados os gráficos presentes na (Figura 9), com o intuito de avaliar a velocidade da digestão dos indivíduos, de tal forma que possa servir como ferramenta para diagnósticos referentes ao mau funcionamento da motilidade gástrica. No eixo x dos gráficos temos cada ponto no tempo (15, 30, 45, 60, 90, 120, 180 e 240) e no eixo y temos o número de componentes principais (NCP) obtidas para explicar 95% da

variabilidade total contida nas imagens.

É possível perceber que, o padrão de comportamento do processo digestivo dos indivíduos 1, 2 e 3 é semelhante, pois ao final das quatro horas de acompanhamento, os mesmos conseguiram digerir todo o alimento ingerido. Além disso, possuem somente um pico no gráfico, representando o ápice da digestão, o período que o indivíduo digeriu uma maior quantidade de conteúdo gástrico. Note ainda, que o indivíduo 1 possui o ápice da digestão na primeira hora. Enquanto, os indivíduos 2 e 3 possuem o ápice da digestão no início da segunda hora.

O indivíduo 4 possui um padrão de comportamento digestivo diferente dos demais, sendo que ao final das quatro horas de acompanhamento, o mesmo não conseguiu digerir completamente o alimento. Além disso, o indivíduo 4 possui dois momentos de maior digestão, representados pelos dois picos no gráfico (d) (Figura 9).

Os indivíduos 5 e 6 possuem um comportamento digestivo semelhante, os mesmos apresentam dois picos, sendo o primeiro na primeira hora e segundo entre os 90 e 120 minutos após a ingestão da refeição-teste. Observa-se ainda, que diferentemente do indivíduo 5, o indivíduo 6 não digeriu completamento o alimento ao final das quatros horas, porém o mesmo reteve menos conteúdo gástrico do que o indivíduo 4, que não digeriu grande parte do alimento.

Percebe-se ainda que os indivíduos 1, 2 e 3, possuem um esvaziamento gástrico mais rápido que os demais indivíduos. Dentre os três, o indivíduo 1 possui a taxa mais rápida de digestão.

Com o intuito de avaliar se seria possível agrupar os indivíduos de acordo com seus padrões de comportamento digestivo, realizou-se um agrupamento hierárquico aglomerativo, em que optou-se em utilizar o método HCLUST (Hierarchical Clustering), que consiste na realização de um agrupamento hierárquico utilizando uma matriz composta pelas dissimilaridades dos elementos, em vez da matriz dos dados. Inicialmente, cada elemento representa um grupo, a cada etapa do processo de agrupamento, os dois grupos mais similares são agrupados. Sendo que esse processo continua até a obtenção de um único grupo, formado por todos os elementos.Vale mencionar que foi utilizada a dissimilaridade média entre os elementos. (Rousseeuw (1990) e Johnson e Wichern (2002))

O resultado obtido com o agrupamento encontra-se na Figura 10. Observa-se quatro grupos, sendo o Grupo 1 composto pelo indivíduo 1, o Grupo 2 com os indivíduos 2 e 3, o Grupo 3 com os indivíduos 5 e 6 e o Grupo 4 com indivíduo 4. Além disso, nota-se que o indivíduo 1 se assemelha mais com os indivíduos 2 e 3 do que com os demais, sendo que este resultado já havia sido observado quando estávamos avaliando a velocidade da digestão dos indivíduos (Figura 9). Enquanto, que o indivíduo 4 se assemelha mais com os indivíduos 5 e 6, por apresentar um processo digestivo mais lento.

Sendo assim, é possível conectar os resultados obtidos no estudo da velocidade do processo digestivo com o agrupamento obtido, podendo observar que os grupos representam quatro taxas de esvaziamento gástrico, sendo o primeiro mais rápido, o segundo moderadamente rápido, o terceiro moderadamente lento e o quarto com o processo digestivo mais lento em relação aos demais.

Com o intuito de verificar se o agrupamento obtido condiz com a realidade dos dados, foram expostas na Figura 11, as imagens representadas pelas regiões de interesse para cada indivíduo em quatro pontos no tempo (15, 60, 120 e 240 minutos). Sendo possível perceber que os indivíduos agrupados pelo Dendograma (Figura 10) possuem um comportamento digestivo semelhante. Portanto, o agrupamento parece está condizente com os dados analisados. Vale ressaltar, que as conclusões feitas neste estudo precisam ser verificadas com um profissional da área, porém não temos evidências para concluir que as mesmas não estão condizentes.

Com os três estudos, foi possível alcançar os objetivos deste trabalho, que consistiam em obter a região de interesse de forma automática por meio da limpeza das imagens e avaliar a velocidade da digestão, com o intuito de servir como ferramenta para diagnósticos de distúrbios gastrointestinais resultantes de um mau funcionamento da motilidade gástrica, ocasionando um esvaziamento gástrico anormal.

Figura 10 – Dendograma

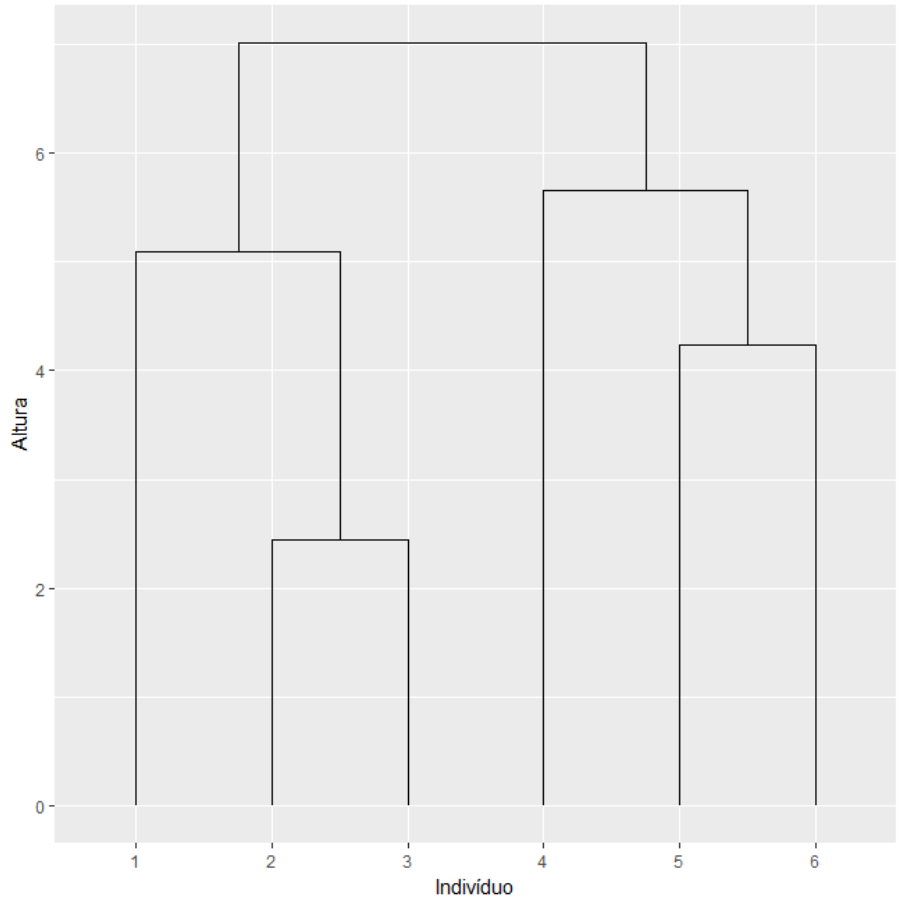
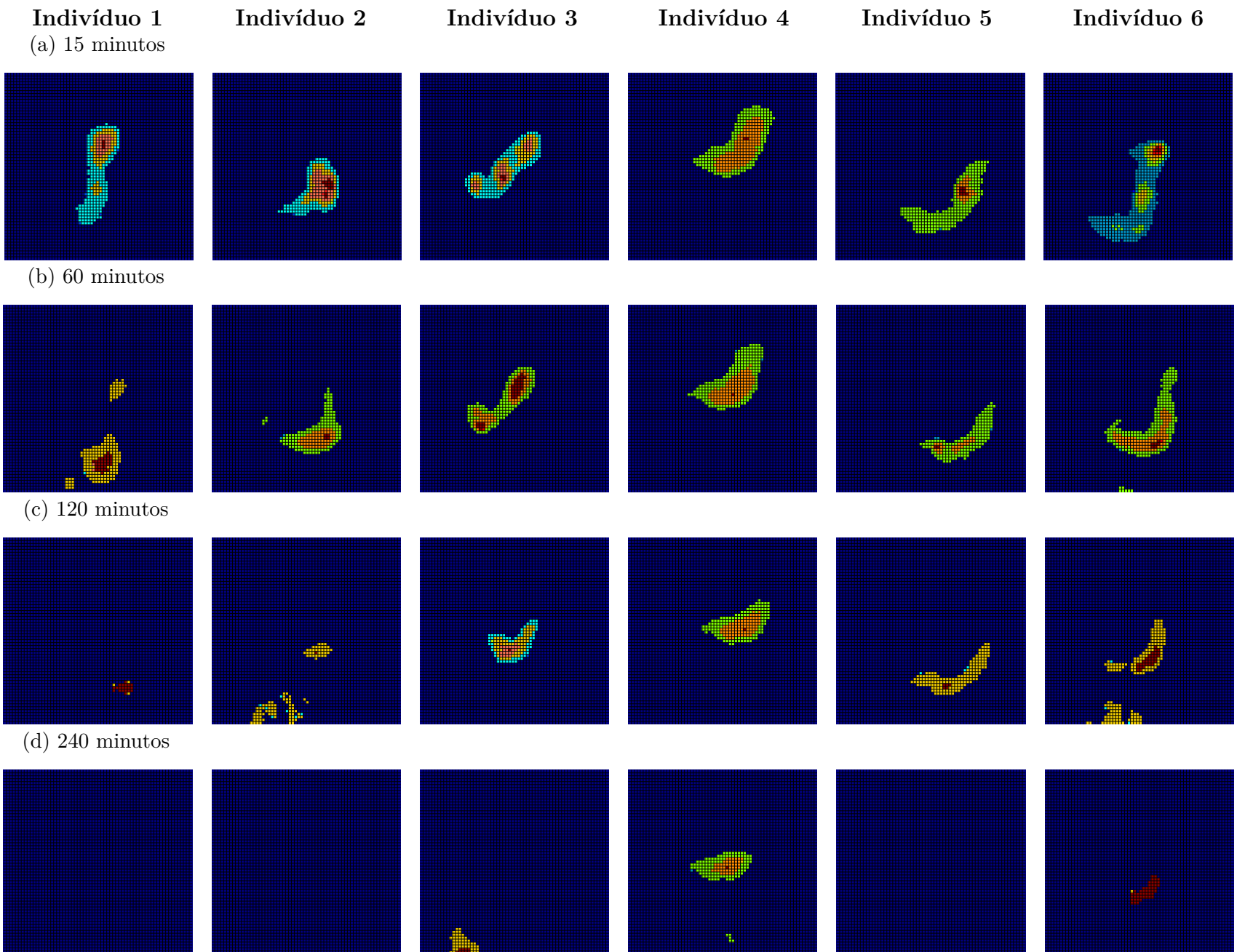


Figura 11 – Comparação dos indivíduos de acordo com o agrupamento obtido



3 Conclusão

Este trabalho consistiu em analisar imagens cintilográficas do estômago de seis indivíduos, com o intuito de avaliar a motilidade gástrica dos mesmos. Tendo como objetivos, a identificação da região de interesse de forma automática e avaliação da velocidade do processo digestivo de cada indivíduo, servindo como ferramenta para identificação de possíveis distúrbios gastrointestinais, tais como Gastroparesia.

Na identificação da região de interesse, subdividiu-se as imagens de cada ponto no tempo em dois grupos, e utilizou-se a mediana das imagens para que fosse possível obter duas imagens representativas para cada ponto no tempo. Com isso, foi possível identificar a região de interesse sem utilizar técnicas para redução da dimensionalidade dos dados, como a Análise de Componentes Principais.

Para avaliar a velocidade do processo digestivo dos indivíduos foi utilizada a Análise de Componentes Principais Bidimensional, em que as imagens foram divididas em blocos e aplicou-se a ACP Tradicional para identificação do comportamento do processo digestivo, em que gráficos foram gerados com o intuito de captar a capacidade de digestão de cada indivíduo durante um período de quatro horas.

Observou-se que quatro dos seis indivíduos conseguiram digerir o alimento no período de acompanhamento. Além disso, foi possível perceber que três indivíduos apresentaram um esvaziamento gástrico mais rápido que os demais. Com tais informações foi possível agrupar os indivíduos com comportamento digestivo semelhante, obtendo quatro grupos, que representasse as diferentes taxas de esvaziamento gástrico, sendo elas: rápida, moderadamente rápida, moderadamente lenta e lenta. Além disso, observou-se que dois grupos foram compostos somente por um indivíduo, sendo eles o indivíduo 1, que apresentou a digestão mais rápida entre os demais (Grupo 1) e o indivíduo 4, com o esvaziamento gástrico mais lento e inabilidade de digerir o alimento nas quatro horas de acompanhamento (Grupo 4).

Foi possível perceber que as técnicas utilizadas neste trabalho são boas opções para avaliação da motilidade gástrica, sendo que atualmente o processo utilizado para realizar este estudo acaba se tornando muitas vezes complicado. Vale ressaltar que as conclusões feitas neste trabalho precisam ser verificadas por um técnico da área.

Vale ressaltar que outras técnicas poderiam ter sido utilizadas para alcance dos objetivos do estudo deste trabalho, ficando para estudos futuros a utilização das mesmas e comparação com os resultados obtidos neste trabalho. Além disso, vale lembrar a existência de um pacote para leitura das imagens diretamente no R (*oro.dicom*) Thornton (2011), porém existem algumas dúvidas acerca dos resultados obtidos pelo mesmo, sendo

interessante futuramente analisar e identificar os possíveis erros presentes no pacote.

Referências

DAVIS, C. S. *Statistical methods for the analysis of repeated measurements*. [S.l.]: Springer, 2002.

DWIVEDI, A. Color image compression using 2-dimensional principal component analysis (2dpca). The 9th Asian Symposium on Information Display, New Delphi, India, 2008. 12

HASTIE, H. Z. T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 2000. 7, 8

HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. [S.l.]: Chapman & Hall/CRC, 2015. 6, 7, 8, 10

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970. 8

JEFFERS, J. N. R. Two case studies in the application of principal component analysis. 1967. 10

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. [S.l.]: Prentice Hall, 2002. 6, 8, 26

MALMUD, L. S. et al. Scintigraphic evaluation of gastric emptying. In: ELSEVIER. *Seminars in nuclear medicine*. [S.l.], 1982. v. 12, n. 2, p. 116–125. 24

ROUSSEEUW, L. K. P. J. *Finding groups in data: An introduction to cluster analysis*. [S.l.]: Wiley Interscience, 1990. 26

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern recognition*. [S.l.]: third ed. Academic Press, 2006. 3

THORNTON, B. W. J. S. A. Working with the DICOM and NIfTI data standards in R. *Journal of Statistical Software*, 2011. Disponível em: <http://www.jstatsoft.org/v44/i06/>. 15, 16, 31

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 1996. 7

YANG, J. et al. Two-dimensional pca: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, 2004. 13

APÊNDICE A – Código R

```
# GERACAO DOS GRAFICOS (TEMPOXCOMP) COM 2 SUBDIVISOES EM CADA TEMPO ----

# FUNCOES  QUE SERAO UTILIZADOS PARA GERACAO DAS IMAGENS E GRAFICOS:
rotate <- function(x) t(apply(x, 2, rev))

library(readxl)
my_palette <- colorRampPalette(c("navy", "blue", "deepskyblue3", "cyan",
                 "chartreuse1", "yellow", "gold2",
                 "orange", "darkorange", "coral1",
                 "red2", "brown1", "darkred"))(n = 10)

#IMAGENS A SEREM IMPORTADAS DE TODOS OS INDIVIDUOS
#LEITURA DAS IMAGENS:

pasta <- list.files('D:/DAS 00X - XLS')
planilhas <- list()

for (i in 1:length(pasta)) {
  planilhas[[i]] <- list.files(paste('D:/DAS 00X - XLS', pasta[i],
                      sep = "/"))
}

imagens <- list()
for (i in 1:length(pasta)) {
  imagens[[i]] <- list()
  for (j in 1:length(planilhas[[i]])) {
    imagens[[i]][[j]] <- as.matrix(rotate(read.csv(paste("D:/DAS 00X -
   XLS",
                  pasta[i], planilhas[[i]][j], sep = "/"), header = F)))
    print(j)
  }
  imagens[[i]] <- do.call(cbind, imagens[[i]])
  dim(imagens[[i]]) <- c(64, 64, 4096)
}


# OS PARAMETROS DA FUNCAO Graficos ABAIXO REPRESENTAM:

# ind = individuo, que assume valores inteiros de 1 a 6,
# 1 = individuo 005
# 2 = individuo 006
# 3 = individuo 007 ...
# 6 = individuo 013
```

```
# pct = limiar da proporcao acumulada explicada pelas componentes

Graficos <- function(ind, pct, sps) {
  library(ggalt)
  library(ggplot2)

  #15 MINUTOS
  med_015_1 <- apply(imagens[[ind]][,,1:256], c(1,2), median)
  med_015_2 <- apply(imagens[[ind]][,,257:512], c(1,2), median)

  #30 MINUTOS
  med_030_1 <- apply(imagens[[ind]][,,513:769], c(1,2), median)
  med_030_2 <- apply(imagens[[ind]][,,770:1024], c(1,2), median)

  #45 MINUTOS
  med_045_1 <- apply(imagens[[ind]][,,1025:1280], c(1,2), median)
  med_045_2 <- apply(imagens[[ind]][,,1281:1536], c(1,2), median)

  #60 MINUTOS
  med_060_1 <- apply(imagens[[ind]][,,1537:1792], c(1,2),median)
  med_060_2 <- apply(imagens[[ind]][,,1793:2048], c(1,2),median)

  #90 MINUTOS
  med_090_1 <- apply(imagens[[ind]][,,2049:2304], c(1,2),median)
  med_090_2 <- apply(imagens[[ind]][,,2305:2560], c(1,2),median)

  #120 MINUTOS
  med_120_1 <- apply(imagens[[ind]][,,2561:2816], c(1,2),median)
  med_120_2 <- apply(imagens[[ind]][,,2817:3072], c(1,2),median)

  #180 MINUTOS
  med_180_1 <- apply(imagens[[ind]][,,3073:3328], c(1,2), median)
  med_180_2 <- apply(imagens[[ind]][,,3329:3584], c(1,2), median)

  #240 MINUTOS
  med_240_1 <- apply(imagens[[ind]][,,3585:3840], c(1,2), median)
  med_240_2 <- apply(imagens[[ind]][,,3841:4096], c(1,2), median)

  img_m_g <- list(med_015_1, med_015_2, med_030_1, med_030_2, med_045_1,
                  med_045_2, med_060_1, med_060_2, med_090_1, med_090_2,
                  med_120_1, med_120_2, med_180_1, med_180_2,
                  med_240_1, med_240_2)

  #DEFININDO AS DEMARCACOES PARA COMPUTAR OS BLOCOS DE TAMANHO L X L
  i <- 1
  j <- 16 #VALOR DE L, NO CASO 16
  l <- c()
```

```
  c <- c()
  while (i <= 64 & j <= 64) {
    l <- c(l,i)
    c <- c(c,j)
    i <- i + 16
    j <- j + 16
  }

  ## DEFININDO OS BLOCOS:
  m <- (64 / 16) ^ 2 # NUMERO DE BLOCOS
  pca <- list()
  blocos <- list()
  # T REPRESENTA O TEMPO, TENDO 8 VALORES POSSIVEIS,
  # SENDO 1 = 15 MIN, 2 = 30, ETC
  for (t in 1:length(img_m_g)) {
    j <- 1
    while (j <= m) {
      # K E i REPRESENTAM AS DEMARCACOES NAS LINHAS E COLUNAS]
      # (RESPECTIVAMENTE) EM QUE A IMAGEM DEVERA SER DIVIDIDA
      for (k in 1:length(l)) {
        for (i in 1:length(c)) {
          blocos[[j]] <- as.matrix(img_m_g[[t]][l[k]:c[k], l[i]:c[i]])
          colnames(blocos[[j]]) <- NULL
          rownames(blocos[[j]]) <- NULL
          j <- j + 1
        }
      }
    }
    # TRANSFORMANDO OS BLOCOS EM VETORES, ARMAZENADO EM UMA MATRIZ: X
    x <- NULL
    for (i in 1:length(blocos)) {
      x <- cbind(x, as.vector(t(blocos[[i]])))
    }

    # COMPUTANDO AS COMPONENTES PRINCIPAIS NA MATRIZ X PARA CADA PONTO
    # NO TEMPO T
    pca[[t]] <- princomp(x)
  }


  # DETERMINANDO O NUMERO DE COMPONENTES NECESSARIAS PARA EXPLICAR UMA
  # PROPORCAO ESPECIFICADA (PCT) DA VARIABILIDADE
  componentes <- c(rep(0, 16))
  p <- 1
  while (p <= 16 && sum(img_m_g[[p]][14:52, 14:52] >= 0.1 &
                img_m_g[[p]][14:52, 14:52] <= max(img_m_g[[p]])) > 0) {
    acumulado <- round(cumsum((pca[[p]]$sdev)^2)/sum(pca[[p]]$sdev^2),2)
```

```
    j <- 1
    comp <- 1
    while (acumulado[j] < pct & j <= 16) {
      j <- j + 1
      comp <- comp + 1
    }
    componentes[p] <- comp
    if (j == 1) {
      componentes[p] <- 1
    }
    p <- p + 1
  }


  tabela <- as.data.frame(c(7.5, 15, 22.5, 30, 37.5, 45,52.5, 60, 75,
                            90, 105, 120, 150, 180, 210, 240))
  tabela$`Numero de Componentes` <- componentes
  rownames(tabela) <- NULL
  colnames(tabela) <- c("Tempo", "Numero de Componentes")

  t_g <- ggplot(tabela, aes(x = Tempo, y = `Numero de Componentes`)) +
    labs(title = paste("NCP versus Tempo -",pct*100,"% var exp"),
       x = "Tempo", y = "NCP") +
    geom_xspline(group = 1, color = "#00B0F6", spline_shape = sps) +
    theme(plot.title = element_text(hjust = 0.5)) +
    scale_x_discrete(limits = c(15, 30, 45, 60, 90, 120, 180, 240)) +
    scale_y_continuous(limits = c(-0.1, 5.1))

  resultados <- list(grafico = t_g, tabela = tabela)

  return(resultados)
}

# IMAGEM MEDIANA AOS 15 MINUTOS
# E NECESSARIO SALVAR A LISTA IMG_M_G COMO OBJETO R,
# ANTES DE EXECUTAR ESTE COMANDO
image(img_m_g[[1]], col = my_palette)
grid(64, 64, lwd = 2, col = "black", lty = 1)

#RESULTADOS PARA TODOS OS INDIVIDUOS
a <- list()
tab.agrup <- data.frame(rep(0, 16), rep(0, 16))
for (i in 1:6) {
  a[[i]] <- Graficos(i, 0.95, 1.5)
  tab.agrup <- cbind(tab.agrup,a[[i]]$tabela)
}
```

```
tab.agrup <- tab.agrup[, -c(1, 2, 5, 7, 9, 11, 13)]
colnames(tab.agrup) <- c("Tempo", "Ind 1", "Ind 2", "Ind 3", "Ind 4",
                         "Ind 5", "Ind 6")

#GERACAO DOS GRAFICOS
pdf("C:/Users/acaro/Dropbox/TCC/TCC - Atualizado/imagens/Graficos.pdf")
for (i in 1:6) {
  plot(a[[i]]$grafico)
}
dev.off()

#GERACAO DO DENDOGRAMA
library("ggdendro")
g <- hclust(dist(t(tab.agrup[, -1])), method = "average")
ggdendrogram(g, rotate = FALSE, size = 2)
ddata  <-  dendro_data(g, type = "rectangle")
p <-  ggplot(segment(ddata )) +
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend)) +
  labs(x = "Individuo", y = "Altura") +
  scale_x_discrete(limits = c(1, 2, 3, 4, 5, 6))
p
```